

INFORMATION RETRIEVAL DEVICE, AND METHOD THEREFOR

Patent number: JP10198706

Publication date: 1998-07-31

Inventor: SAKATA TAKESHI

Applicant: DIGITAL VISION LAB KK

Classification:

- international: G06F17/30; G06F9/44

- european:

Application number: JP19970313766 19971114

Priority number(s): JP19960304840 19961115; JP19970313766 19971114;
US19970970625 19971114

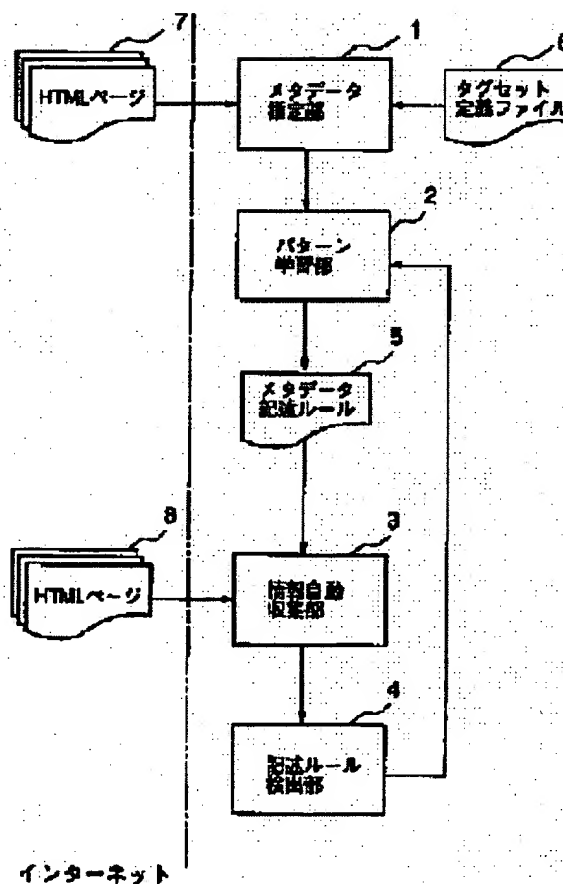
Also published as:

 US6085190 (A1)

Report a data error here

Abstract of JP10198706

PROBLEM TO BE SOLVED: To enable a consumer to quickly and easily retrieve desired commodities, by preparing a pattern learning means, which produces a rule to extract the information having a designated attribute based on this attribute. **SOLUTION:** A meta-data designation part 1 refers to a tag set definition file 6 to designate the meta-data on an HTML page 7 such as a prescribed shopping mail, etc., and to output the meta-data to a pattern learning part 2. The part 2 inputs the designated meta-data and learns the description pattern of the meta-data to produce a meta-data description rule 5. An automatic information collecting part 3 collects the meta-data from an HTML page 8 based on the rule 5. A description rule verification part 4 verifies the validity of the rule 5 based on the meta-data collected from the page 8 and outputs the result of this verification to the part 2. The part 2 updates the rule 5 based on the received result of verification.



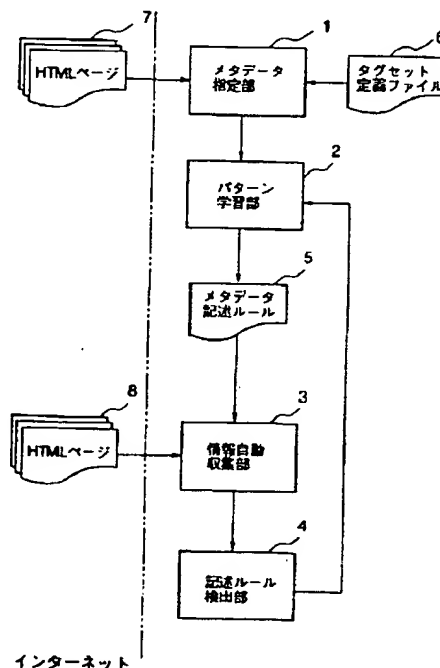
Data supplied from the esp@cenet database - Worldwide

(11)特許出願公開番号

(43)公開日 平成10年(1998)7月31日

審査請求 有 請求項の数12 OL (全 10 頁)

(74)代理人 弁理士 鈴江 武彦 (外5名)



【特許請求の範囲】

【請求項1】 様々な記述形式で記載された情報の少なくとも1つの属性を指定するメタデータ指定手段と、前記指定された属性に基づいて、その属性を有する情報を抽出するためのルールを作成するパターン学習手段と、

を具備することを特徴とする情報検索装置。

【請求項2】 前記様々な記述形式で記載された情報はネットワークで接続された複数のデータベースに格納された情報であることを特徴とする請求項1記載の情報検索装置。

【請求項3】 前記ルールは、前記指定された属性情報の抽象化によって作成されることを特徴とする請求項1記載の情報検索装置。

【請求項4】 前記ルールに基づいて前記指定された属性を有する情報を収集する情報収集手段を更に具備することを特徴とする請求項1記載の情報検索装置。

【請求項5】 前記ルールに基づいて収集された情報に所望以外の情報が含まれているかどうか検証する検証手段を更に具備することを特徴とする請求項4記載の情報検索装置。

【請求項6】 前記パターン学習手段又は前記検証手段のいずれかが前記検証手段の検証結果に基づいてルールの更新処理を行うことを特徴とする請求項5記載の情報検索装置。

【請求項7】 前記更新処理は前記属性情報の具体化処理を含むことを特徴とする請求項6記載の情報検索装置。

【請求項8】 様々な記述形式で記載された情報の少なくとも1つの属性を指定し、前記指定された属性に基づいて、その属性を有する情報を抽出するためのルールを作成することを特徴とする情報検索方法。

【請求項9】 前記ルールは、一致する文字列はそのまま残し、一致しない文字列を変数に変換する抽象化処理によって作成することを特徴とする請求項8記載の情報検索方法。

【請求項10】 前記ルールに基づいて前記指定された属性を有する情報を収集することを特徴とする請求項8記載の情報検索方法。

【請求項11】 前記ルールに基づいて収集された情報に所望以外の情報が含まれているかどうか検証することを特徴とする請求項10記載の情報検索方法。

【請求項12】 前記検証結果に基づいてルールの更新処理を行うことを特徴とする請求項11記載の情報検索方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、様々な記述形式でデータベース化された例えば商品情報の情報検索装置及

びその方法に関し、特に、例えば、インターネットのような通信における異なる提供者によって提供された種々の商品情報に対する情報検索装置及びその方法に関する。

【0002】

【従来の技術】 近年、パソコン通信或いはインターネットなどによる仮想ショッピングモール或いはショッピングページを使用した通信販売が脚光を浴びている。しかし、この仮想ショッピングモール或いはショッピングページにおいて、商品を購入する消費者としては、探したい物が見つからない等の問題を有している。また、商品を提供する提供者としては、客が店に来てくれない（又は、ホームページにアクセスしてくれない）という問題を有している。ここで、商品とは、有形の商品ばかりではなく、無形の商品を含み、例えば、商品の提供者が放送事業者の場合には、商品とは放送される番組のようなサービスをいうものとする。

【0003】 上記の問題において、消費者側の問題点として、探したい商品が見つからないとは以下のような状態を指している。放送番組において、聴きたい曲を放送している番組がわからないとか、ある役者が出ている映画を見たいのだが、番組表では大まかなことしか書いてないので、どの映画に出ているのかわからないといった状態を指し、また、例えば、インターネットにおいて、ある商品を買っているホームページを探そうとした場合に、どのホームページを見ればその商品を探せるのかわからないといった状態を指している。また、上記のように、具体的に探したい商品が定まっておらず、何かおもしろい番組はないか、最近評判の商品は何かといったように、探したい物が具体的にわからない場合もある。

【0004】 更に、提供者の側から見た場合には、現在のインターネットの検索サービスを見ると消費者側から商品をアプローチするシステムをとっているために、インターネット上に新しいWWWのサイトを開いても消費者に見つけてもらえない。

【0005】 また、インターネットの検索サービスは、全文検索技術を使っており、キーワードによる検索しかできない。そのため、赤いポロシャツが欲しい場合でも、赤とポロシャツという2つのキーワードの論理積で検索することになり、同一ページに赤いTシャツと黄色のポロシャツが販売されているページか、該当結果として帰ってきてしまい、検索結果にユーザの意図した物とは異なる物が入ってしまうという問題点がある。また、従来の検索サービスは、キーワードを文字列としか扱わないため、例えば、5000円以下という数値の範囲指定ができない。

【0006】

【発明が解決しようとする課題】 上記のように、現在の商品検索においては、消費者側としては、探したい商品が見つからない、更には、キーワードによる全文検索に

なるので、検索結果に所望の物と異なる物が多く含まれてしまうといった問題があり、提供者としては、客が店に来てくれないといった問題がある。

【0007】本発明は、上記の課題を考慮してなされたもので、その目的とするところは、消費者が所望の商品を迅速かつ容易に検索でき、その結果として、提供者が特別な努力をすることなく消費者に提供できる商品を消費者に提示できる情報検索装置及びその方法を提供することである。

【0008】

【課題を解決するための手段】本発明は、上記の課題を解決するために次のような手段を講じた。本発明の情報検索装置は、様々な記述形式で記載された情報の少なくとも1つの属性を指定するメタデータ指定手段と、前記指定された属性に基づいて、その属性を有する情報を抽出するためのルールを作成するパターン学習手段とを備えたことを特徴とする。

【0009】上記の構成において、以下のような実施態様を有することが好ましい。

(1) 前記様々な記述形式で記載された情報はネットワークで接続された複数のデータベースに格納された情報であること。

(2) 前記ルールは、前記指定された属性情報の抽象化によって作成されること。

(3) 前記ルールに基づいて前記指定された属性を有する情報を収集する情報収集手段を更に備えること。

(4) 前記ルールに基づいて収集された情報に所望以外の情報が含まれているかどうか検証する検証手段を更に備えること。

(5) 前記パターン学習手段又は前記検証手段のいずれかが前記検証手段の検証結果に基づいてルールの更新処理を行うこと。

(6) 前記更新処理は前記属性情報の具体化処理を含むこと。

【0010】本発明の情報検索方法は、様々な記述形式で記載された情報の少なくとも1つの属性を指定し、前記指定された属性に基づいて、その属性を有する情報を抽出するためのルールを作成することを特徴とする。本情報検索方法の好ましい実施態様は以下の通りである。

(1) 前記ルールに基づいて前記指定された属性を有する情報を収集すること。

(2) 前記ルールに基づいて収集された情報に所望以外の情報が含まれているかどうか検証すること。

(3) 前記検証結果に基づいてルールの更新処理を行うこと。

【0011】上記のような構成により、消費者にとっては、所望の商品情報を迅速かつ容易に抽出することが可能であるとともに、所望の商品情報を抽出するためのルールの作成・更新も非常に容易な商品情報の検索が可能となる。更に、情報提供者においては、ルール化された

メタデータを用いて情報自動収集手段が該当する情報を収集し、それを用いて情報の正確な案内が行われるので、HTMLページを消費者が見てくれないという不都合がなくなる。

【0012】

【発明の実施の形態】図面を参照して本発明の実施の形態を説明する。図1は、本発明の一実施形態に係る情報検索装置の概略構成を示すブロック図である。以下、インターネットにおけるWWWサイトのホームページ（以下、HTML（HyperText Markup Language）ページと称する）による商品検索について説明するが、本発明はそれに限らず、異なるデータベースを有するようなネットワーク（インターネット、イントラネットを含む）上の商品検索にも適用可能である。

【0013】本実施形態に係る情報検索装置は、メタデータ指定部1と、パターン学習部2と、情報自動収集部3と、記述ルール検証部4とからなる。メタデータ指定部1は、タグセット定義ファイル6を参照して、所定のショッピングモールなどのHTMLページ7からメタデータを指定し、パターン学習部1に指定したメタデータを出力する。ここで、メタデータとは、物（例えば、商品）の意味を表す情報であって、その属性と属性値（例えば、属性を「価格」とし、属性値を実際の価格とする）を含むデータである。

【0014】パターン学習部2は、詳細は後述するように、メタデータ指定部1で指定されたメタデータを入力して、メタデータの記述パターンを学習して、メタデータ記述ルール5を作成する。

【0015】情報自動収集部3は、メタデータ記述ルール5に基づいてHTMLページ8からメタデータを収集する。記述ルール検証部4は、収集したHTMLページ8のメタデータに基づいてメタデータ記述ルール5の妥当性を検証し、その結果をパターン学習部1に出力する。パターン学習部1は、前記結果に基づいて、メタデータ記述ルール5を更新する。

【0016】上記のように構成された情報検索装置の詳細を説明する。例えば、商品カタログにおいて、図2に示すような「ファッション：婦人服」のHTMLページを指定したものとする。この場合には、タートルネックセーター及びベロアドルマンTシャツの2つの商品が表示されている。これらの商品の属性として、(商)品名、価格、送料、素材等が記載されている。図2に示すようなHTMLページの一部について、その記載例を図3に示す。図3において、<>で囲まれた部分がタグと呼ばれるもので、表示データの種類、位置等を示す情報が記載されている。

【0017】図2に示した商品を例にとると、指定者はメタデータ指定部1において、小片名称という属性のメタデータとして、「タートルネックセーター」と、「ベロアドルマンTシャツ」を指定する。メタデータ指定部

1は、図4(a)及び(b)に示すように、指定された値の前後のHTMLテキストを一定の長さだけ付加して、パターン学習部2に渡す。

【0018】この2つのデータからパターン学習部2は、図4(c)に示すように、一致する部分はそのままの文字列として残し、一致しない部分を他の文字列に置き換えるような処理を行う。例えば、図4(a)と(b)とのHTML表示において、「品名」と「商品名」、「2085-26907」と「2086-26918」、「タートルネックセーター」と「ペロアドルマンTシャツ」が異なっている。そこで、図4(c)のように、一致しない部分の文字列を、単に文字列が存在するという意味の「text」で置き換える(この文字列を以下、「ワイルド・カード」と称する。また、図4(c)では、指定された属性の値が記述されている場所は「*****」で表している)。すなわち、パターン学習部2は、図4(c)に示すように、同一の検索対象において、異なる文字列で記載された部分はワイルド・カードで文字列を置き換えることによって、ルールを作成する。そして、同一の検索対象において、更に異なる記述部分がある場合には、その異なる部分をワイルドカードに置き換えて、異なる文字列の記載があった場合であっても検索可能になるようなルールを作成していく。このような、特定の文字列を順次ワイルド・カードに置き換えていく処理を「抽象化処理」と称し、逆に、ワイルド・カードを実際の文字列に置き換えていく処理を「具体化処理」と称する。

【0019】パターン学習部2による抽象化処理の方法について、図05を参照して説明する。図05は、抽象化の例を示す階層図である。図5にある $(X \wedge a \wedge b)$ とは、1つの照合ルールを表しており、Xというタグ名にaという属性とbという属性がついたHTMLタグとこのルールがマッチすることを表している。記号 \wedge は条件がANDで結ばれていることを意味する。()は、()内の条件が1つのHTMLタグと照合することを意味する。また、aという属性に特定の値が結びついていることが条件の場合をa0、値の内容は問わずaという属性があればよい場合は、a1とする。第1層の $(X \wedge a \wedge b) \wedge (y \wedge c)$ とは、HTML文書中に定められた値を持つaという属性とbという属性を持つXタグがまず存在し、その直後に定められた値を持つcという属性を持つYタグが存在するときに、台致するという意味の照合ルールである。

【0020】次に、照合ルールの抽象化について考える。照合の条件としては、a1の方がa0よりも条件的に緩く抽象化されているといえる。a1よりも抽象化する場合は、属性aはあってもなくても良いということになり、ルールからはa1は消去される。抽象化は、()単位で行われ、()の中の属性に関する条件がまず抽象化され、照合ルールがXなどタグ名のみとなった場合は、次に抽象化するときはそのHTMLタグがあっても

なくても良いことになり、Xという条件が消去される。【0021】但し、消去してはまずい場合、例えば、 $X \wedge Y \wedge Z$ のようにXとYとZの3つのタグが並んでいることが条件の照合ルールでYを抽象化する場合、Yをただ消去して $X \wedge Z$ とすることはできない(そうすると、 $X \wedge Y \wedge Z$ に台致していたHTML文書が照合ルールに台致しなくなり抽象化とはならない)。この場合は、XとZとの間にどのようなタグでも構わないか、タグが1つ存在することが条件となる。図5では、この条件をTで表している(ここで、Tはタグのワイルドカードに相当する)。先程の $X \wedge Y \wedge Z$ のYの抽象化では $X \wedge T \wedge Z$ となる。また、図5では、これ以上抽象化できないことを*で表現している。また、図5では、説明の単純化のため、タグに含まれない文字列の扱いを省いているが、特定の文字列をsと表して照合ルールの()の外におくことで対応できる。

【0022】上記のような具体化処理により、パターン学習部2は、メタデータのルールを作成し、メタデータ記述ルール5に新たなルールを追加したり、現在のルールを更新したりする。

【0023】実際に、パターン学習部2によって、ルールが作成・更新された後に、情報自動収集部3は、実際にインターネットにアクセスして、HTMLページ7における所望の商品情報を検索する。

【0024】しかし、例えば、図4(c)に示す「商品名称」を探すルールは、品名及び商品名を満たすように「text」(ワイルド・カード)に置き換えられているので、図6に示した価格を表している部分にもマッチングし、「商品名」=「5900円」という間違っただけの抽出を行ってしまう。このような場合には、記述ルール検証部4は、この検索に係るルールが不適切であるとして、パターン学習部2にその旨を出力する。

【0025】このような事態は、ルールの記述を抽象化しすぎたために起こるので、パターン学習部2は、記述ルール検証部4の出力を受けて、抽象化処理の逆の処理である具体化処理により、図4(c)において、ワイルド・カードとして「text」としていた部分のうち、最初の「text」の部分それぞれ、図7(a)及び(b)に示すような「品名」と「商品名」とにしたルールを作成し、図4(c)のルールを更新する。このようにすることにより、図6に示す「価格(税込)」の部分は、品名でも商品名でもないの、検索対象からはずれることになる。

【0026】上記のように、本発明では、HTMLページに基づいて所望の商品情報を抽出するためのルールを抽象化処理によって作成し、そのルールに基づいてHTMLページから所望の商品情報を抽出することができる。更に、ルールの記述が抽象的すぎて所望以外の商品情報も抽出する場合には、具体化処理により、商品情報の絞り込みが可能である。

【0027】従って、本発明によれば、所望の商品情報を迅速かつ容易に抽出することが可能であるとともに、所望の商品情報を抽出するためのルールの作成・更新も非常に容易な商品情報の検索が可能となる。更に、情報提供者においては、記述ルールによって抽出されたデータに基づいて消費者が商品情報を検索できるので、消費者が望む商品を提供しているのに消費者が気付かないという不具合がなくなる。

【0028】本発明は、上記の発明の実施の形態に限定されない。上記実施形態においては、メタデータ指定部1から記述ルール検証部4までを1つのシステムとして記載したが、メタデータ指定部1とパターン学習部2とを含む第1のシステムと、情報自動収集部3からなる第2のシステムと、メタデータ指定部1とパターン学習部2と記述ルール検証部4とを含む第3のシステムとがそれぞれ別々のシステムであっても良い。この場合には、第1から第3のシステムは、それぞれ、メタデータ記述ルール5を含む。なお、これ以外にも、構成の変更が可能である。

【0029】また、ルールの更新を上記の実施形態では、パターン学習部2で行っているが、記述ルール検証部4で行っても良い。また、ルールの抽象化処理或いは具体化処理方法についても、上記の方法に限定されず、具体化及び抽象化が可能であれば、他の種々の方法が適用可能である。

【0030】更に、上記の実施形態においては、インターネット上の商品情報の検索について、説明したが、それに限らず、商品情報ばかりでなく、他の情報についても適用可能であり、データベースの構造が異なるような情報を一括して検索するような情報検索システムにも適

* 用可能である。その他、本発明の要旨を変更しない範囲で種々変形して実施できるのは勿論である。

【0031】

【発明の効果】本発明によれば次のような効果が得られる。上記のように、本発明によれば、消費者にとっては、所望の商品情報を迅速かつ容易に抽出することが可能であるとともに、所望の商品情報を抽出するためのルールの作成・更新も非常に容易な商品情報の検索が可能となる。更に、情報提供者においては、ルール化されたメタデータを用いて情報自動収集部3が該当する情報を収集するので、HTMLページを消費者が見てくれないという不都合がなくなる。

【図面の簡単な説明】

【図1】 本発明の一実施形態に係る情報検索装置の概略構成を示すブロック図。

【図2】 HTMLページの表示例を示す図。

【図3】 HTMLページの記載例を示す図。

【図4】 本発明による抽象化処理の具体例を示す図。

【図5】 本発明の抽象化処理の階層を示す図。

【図6】 図4(c)のルールで誤って検索された部分の記載例を示す図。

【図7】 本発明による具体化処理の具体例を示す図。

【符号の説明】

- 1…メタデータ指定部、
- 2…パターン学習部、
- 3…情報自動収集部、
- 4…記述ルール検証部、
- 5…メタデータ記述ルール、
- 6…タグセット定義ファイル、
- 7、8…HTMLページ。

【図6】

<TR>

<TD ALIGN=RIGHT NOWRAP>価格(税込): </TD>

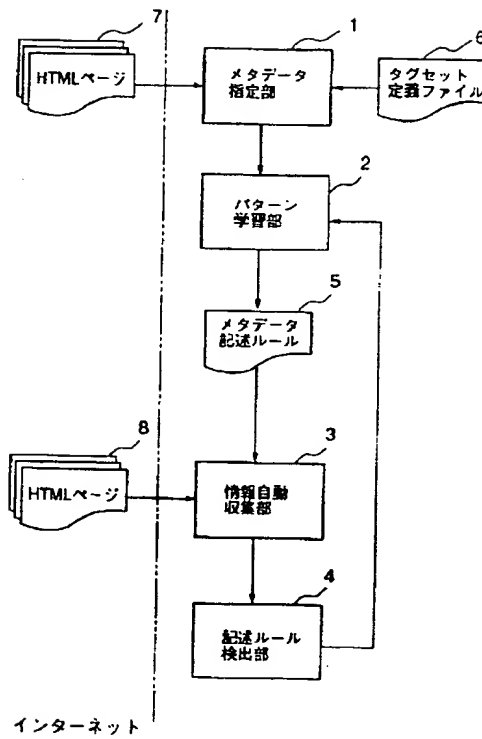
<TD><INPUT NAME="2086-26907:price" TYPE="hidden" VALUE="5,900円">5,900円</TD>

</TR>

<TR>

<TD ALIGN=RIGHT NOWRAP>送料: </TD>

【図1】



【図2】

商品カタログ

！商品ジャンル別！ジャンル別！商品の仕方！商品決定！

ファッション：婦人服

タートルネックセーター

品名	タートルネックセーター		
価格（税込）	5,900円		
送料	300円		
素材	毛100%（カールマーク付）		
洗い方	手洗：X	洗濯機：X	ドライ：O
色	01 ブルー		
サイズ（バスト）	02 M/102（79-87cm） 袖丈51cm		
数量	1		

ペロアドルマンTシャツ

品名	ペロアドルマンTシャツ		
価格（税込）	5,900円		
送料	300円		
素材	ポリエステル100%		
洗い方	手洗：O	洗濯機：O	ネット使用
色	01 ワイン		
サイズ（バスト）	02 M/79-87cm 袖丈70cm		
数量	1		

着心地よく、動きやすく、お洒落心を満足させるプライベートウェア

【図3】

```

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 3.2/EN">
<HTML>
<HEAD>
  <TITLE>商品カタログ:ファッション:婦人服 1</TITLE>
  <META NAME="GENERATOR" CONTENT="Mozilla/3.01b1Gold (Win95; 1) [Netscape]">
</HEAD>
<BODY BGCOLOR="#FFFFFF">

<TABLE WIDTH="100%">
<TR>
<TD ALIGN=RIGHT><1><FONT SIZE=+2>商品カタログ</FONT></1></TD>
</TR>
</TABLE>

<CENTER><P>
<HR> | <A HREF="http://www.commerce.or.jp/cgi-bin/shopping/makeindex">商品ジャンル別</A>
| <A HREF="http://www.commerce.or.jp/cgi-bin/shopping/makeshop">ショップ別</A>
| <A HREF="http://www.commerce.or.jp/cgi-bin/shopping/HowToBuy">買物の仕方</A>
| <A HREF="http://www.commerce.or.jp/cgi-bin/shopping/makesearch">商品検索</A>
|
<HR></P></CENTER>

<TABLE WIDTH="100%">
<TR>
<TD>
<H2><1>ファッション:婦人服 </1></H2>
</TD>

<TD ALIGN=RIGHT><BR>
</TD>
</TR>
</TABLE>

<TABLE WIDTH="100%">
<TR>
<TD ALIGN=LEFT WIDTH=33%><BR>
</TD>

<TD ALIGN=RIGHT WIDTH=33%><BR>
</TD>

<TD ALIGN=RIGHT WIDTH=33%><BR>
</TD>
</TR>
</TABLE>

<P><FORM ACTION="/cgi-bin/shopping/order.cgi" METHOD="POST">
<BR SIZE=3 NOSHADE><INPUT TYPE=HIDDEN NAME=company VALUE="DVL"><INPUT TYPE=HIDDEN NAME=souryou VALUE=
"300"><INPUT TYPE=HIDDEN NAME=keyword VALUE="ファッション:婦人服"></P>

<TABLE BORDER=1 CELSPACING=0 WIDTH="100%">
<TR>
<TD COLSPAN=2>
<TABLE WIDTH="100%">
<TR>
<TD ALIGN=LEFT></TD>

<TD ALIGN=RIGHT></TD>
</TR>

```


【図4】

```

品名:
</TD>
<TD>
<INPUT NAME="2086-26907:title" TYPE="hidden" VALUE="タートルネックセーター">
タートルネックセーター
</TD>
</TR>
<TR>
<TD ALIGN=RIGHT NOWRAP>

```

(a)

```

商品名:
</TD>
<TD>
<INPUT NAME="2086-26918:title" TYPE="hidden" VALUE="ペロアドルマンTシャツ">
ペロアドルマンTシャツ
</TD>
</TR>
<TR>
<TD ALIGN=RIGHT NOWRAP>

```

(b)

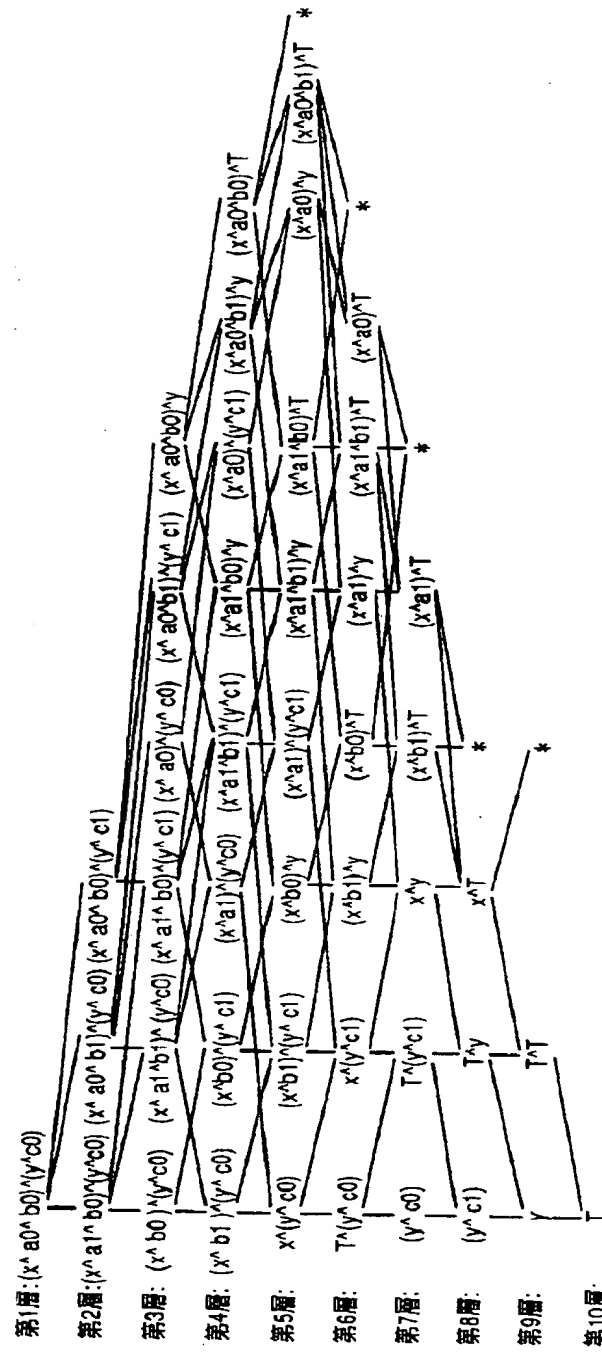
```

text
</TD>
<TD>
<INPUT NAME=text TYPE="hidden" VALUE=text>
*****
</TD>
</TR>
<TR>
<TD ALIGN=RIGHT NOWRAP>

```

(c)

【図5】



【図7】

品名:
 </TD>
 <TD>
 <INPUT NAME=text TYPE="hidden" VALUE=text>

</TD>
 </TR>
 <TR>
 <TD ALIGN=RIGHT NOWRAP>

(a)

商品名:
 </TD>
 <TD>
 <INPUT NAME=text TYPE="hidden" VALUE=text>

</TD>
 </TR>
 <TR>
 <TD ALIGN=RIGHT NOWRAP>

(b)